ESTIMATING THE CONDITIONAL PROBABILITY OF CORRECT CLASSIFICATION USING DEPENDENT TRAINING SAMPLES*

By

S. Bandyopadhyay

Technical Report No. 228

June, 1974

University of Minnesota
Minneapolis, Minnesota

* Supported by U. S. Army Grant DA-ARO-D-31-124-70-G102 at the University of Minnesota, Minneapolis, Minnesota.

ABSTRACT

Five different estimators of the conditional probability of correct classification are considered for the two population classification problem with known covariance matrix where the population distributions follow a stationary Gaussian process. The estimators and their conditional distributions are observed to be identical to the corresponding results for the two population equal sample standard classification problem.

1. Introduction.

In a recent study the author [3] has considered the problem of classifying a unit to one of two specified populations π_1 and π_2. However, unlike the standard classification problem where π_1 and π_2 are considered to be two independent populations, the two papulations are considered as the same population observed at two different states or points of time. Since it is not known to which of the two populations the unit belongs, a vector X of p measurements is used for classification. To be more specific, consider the following set-up.

Let ω be an experimental unit which is a random outcome from a population π. It is known that π is identical to one of the two specified populations π_1 and π_2, where π_1 and π_2 denote the same population at two different points of time t_1 and t_2. Let $X \equiv X(\omega)$ be a $p \times 1$ vector of measurements on the unit ω and the distribution function of X be F_i when $\pi = \pi_i$, $i = 1,2$. The problem is to identify π with one of π_1 and π_2 on the basis of X and the knowledge of F_1 and F_2. When F_1 and F_2 are not completely known, information is obtained about them based on a random sample of N units $\omega_1, \omega_2, \ldots, \omega_N$ with $X_{i\alpha}$ as the X-observation on the unit ω_α observed at time t_i, $\alpha = 1,2,\ldots,N$; $i = 1,2$. Thus the two samples are likely to be dependent since they are based on the same set of units. Samples of such kind will be termed as "dependent training samples". Here $\{X_{i\alpha}, \alpha = 1,2,\ldots,N; \ i = 1,2\}$ constitutes our dependent training sample.

Consider a situation when the distribution of X from π_i, is $N_p[\mu, \Sigma]$, $i = 1,2$; and $(X'_{1\alpha}, X'_{2\alpha})'$, $\alpha = 1,2,\ldots,N$ are independently distributed as

$$(1) \qquad N_{2p}\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma & \rho\Sigma\\ \rho\Sigma & \Sigma\end{pmatrix}\right]$$

where μ_i's are $p \times 1$ unknown vectors and Σ is a known $p \times p$ positive definite matrix and $|\rho| < 1$. To give an example of a situation where the distribution given in (1) arises, consider a first order autoregressive process

$$X_t = m_t + \lambda X_{t-1} + U_t, \quad t = 0, \pm 1, \pm 2, \ldots$$

where U_t's are independently distributed as $N_p[0, \Lambda]$. This process has been studied from the time series point of view for estimation and prediction; it is known that (Anderson [2], p. 166) for every t and r and for $|\lambda| < 1$ and $\sum_{\ell=0}^{\infty} |m_\ell \lambda^\ell| < \infty$

$$\begin{pmatrix}X_{t-r}\\ X_t\end{pmatrix} \sim N_{2p}\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma & \rho\Sigma\\ \rho\Sigma & \Sigma\end{pmatrix}\right]$$

where

$$\mu_1 = \sum_{\ell=0}^{\infty} \lambda^\ell m_{t-r-\ell}, \quad \mu_2 = \sum_{\ell=0}^{\infty} \lambda^\ell m_{t-\ell},$$

$$\Sigma = (1-\lambda)^{-1} \Lambda$$

and

$$\rho = \lambda^r .$$

π_1 and π_2 may be considered as any two time points $t - r$ and t. Under the distribution assumption (1), five estimators of the conditional probability of correct classification (PCC) of a likelihood ratio rule

has been considered in Section 3. Results of Section 3 are generalized for a more general set-up when the training sample consists of observations from a stationary Gaussian process.

2. Classification rules.

Let $f(x|\mu_i)$ be the density frunction of X from π_i, $i = 1,2$. Then the class of rules which classify X to π_1 if, and only if,

$$f(x|\mu_1)/f(x|\mu_2) > c \tag{2}$$

for some constant c is a complete class of admissible rules (Anderson [1], chapter 6). The class of rules given by (2) is equivalent to the class of rules ψ_k^* which classify X to π_1 if, and only if,

$$((X - \mu_1; \Sigma)) - ((X - \mu_2 ; \Sigma)) < k \tag{3}$$

for some constant k and where, for a $p \times 1$ vector y and a $p \times p$ non-singular matrix B,

$$((y)) = y'y \tag{4}$$

and

$$((y ; B)) = y'B^{-1}y \ . \tag{5}$$

k can be taken to be zero when the prior probabilities of drawing X from π_1 and π_2 are the same and misclassification costs are equal, i.e., ψ_0^* is the unique minimax rule in this case. Let $P_i(\psi)$ denote the PCC of the rule ψ given that X is from π_i, $i = 1,2$. Then

$$P_1(\psi_0^*) = P_2(\psi_0^*) = \Phi(\Delta/2) \tag{6}$$

where

$$\Delta^2 = ((\mu_1 - \mu_2 \; ; \Sigma))$$

and $\Phi(x)$ is the distribution function of the standard normal variate. Since ψ_k^* is the best class of rules one can have and since μ_1 and μ_2 are unknowns, one might replace μ_1 and μ_2 by their 'good' estimators to get a plug-in version of the class of rules ψ_k^* given by (3), and a natural choice of a good estimator of μ_i is

$$(7) \qquad \bar{X}_i = (1/N) \sum_{\alpha=1}^{N} X_{i\alpha} , \quad i = 1,2 .$$

Hence, we get a class of rules ψ_k which classifies X to π_1 if, and only if,

$$(8) \qquad ((\bar{X}_1 - X \; ; \Sigma)) - ((\bar{X}_2 - X \; ; \Sigma)) < k .$$

We also note that the rule ψ_k given by (8) is also a minimum distance rule (see Ellison [4]). To judge the performance of the class of rules ψ_k one may compare the PCC of ψ_k with the PCC of ψ_k^*.
Since we are not considering any other rule, we give a justification for using ψ_k. The likelihood of the training sample $\{X_{i\alpha}, \alpha = 1,2,\ldots,N; i = 1,2\}$ and X, when X is from π_1, is given by

$$(9) \qquad L_1 (\mu_1, \mu_2, \rho)$$

$$= c \cdot \exp[- (1/2) \sum_{\alpha=1}^{N} ((X_{2\alpha} - \mu_2 - \rho(X_{1\alpha} - \mu_1); (1-\rho^2) \Sigma))]$$

$$\cdot \exp[- (1/2) \sum_{\alpha=1}^{N} ((X_{1\alpha} - \mu_1; \Sigma)) - (1/2)((X - \mu_1; \Sigma))]$$

where

$$(10) \qquad c = (2\pi)^{-p(2N+1)/2} (1-\rho^2)^{-N/2} |\Sigma|^{-(2N+1)/2} .$$

The supremum of $L_1(\mu_1, \mu_2, \rho)$ over μ_1 and μ_2 is

$$(11) \qquad L_1(\rho) = c \cdot \exp\{\text{trace } [-(1/2)\Sigma^{-1}(A + D_{01})]\}$$

where

$$(12) \qquad S_{ij} = \sum_{\alpha=1}^{N} (X_{i\alpha} - \bar{X}_i)(X_{j\alpha} - \bar{X}_j)', \quad i = 1,2, \quad j = 1,2$$

$$(13) \qquad A = (1-\rho^2)^{-1} (S_{11} - \rho S_{12} - \rho S_{21} + S_{22})$$

and

$$(14) \qquad D_{0i} = [(N/(N+1)](X - \bar{X}_i)(X - \bar{X}_i)', \quad i = 1,2.$$

Similarly, when X is from π_2, $L_2(\rho)$, the supremum of the likelihood over μ_1 and μ_2 is the same as the one given by (11) with D_{02} instead of D_{01} and hence the likelihood ratio rule is obtained by taking ratio, which in equivalent to the rule ψ_k. We also note that ψ_k is still the likelihood ratio rule even if ρ were unknown since

$$L_1(\rho)/L_2(\rho) \equiv \sup_{\rho} L_1(\rho) / \sup_{\rho} L_2(\rho) .$$

It has also been proved that (see [3]) ψ_k is an admissible Bayes rule. We shall take $k = 0$ in the sequel, which is equivalent to assuming equal prior probabilities and equal cost of misclassification. We wish to estimate the PCC of ψ_0 given the training sample, or equivalently, given $\bar{X}_1$ and $\bar{X}_2$ since $\bar{X}_1$ and $\bar{X}_2$ are jointly sufficient (and

complete) for μ_1 and μ_2. Let $P_i(\psi_0|\bar{X}_1, \bar{X}_2)$ denote the conditional PCC of the rule ψ_0 given $\bar{X}_1$ and $\bar{X}_2$ when X is from π_i, i = 1,2. Remarks have been made on the significance of the conditional PCC $P_i(\psi_0|\bar{X}_1, \bar{X}_2)$ and the unconditional PCC $P_i(\psi_0)$ in Hills [7] and Sorum [14] and [15].

3. Estimation of $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$.

Since Σ is known, without loss of generality consider $\Sigma = I_p$.

Now

$$(15) \qquad P_1(\psi_0|\bar{X}_1, \bar{X}_2) = P_1[((\bar{X}_1 - X)) - ((\bar{X}_2 - X)) < 0|\bar{X}_1, \bar{X}_2]$$

$$= P_1[(2X - \bar{X}_1 - \bar{X}_2)'(\bar{X}_2 - \bar{X}_1) < 0|\bar{X}_1, \bar{X}_2]$$

$$= \Phi\,[(\bar{X}_1 + \bar{X}_2 - 2\mu_1)'(\bar{X}_2 - \bar{X}_1)/2d]$$

where

$$(16) \qquad d^2 = ((\bar{X}_1 - \bar{X}_2)) \,.$$

Various estimators of $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$ have been studied by Sorum [14] and [15] of which a few will be considered here.

(a) Plug-in estimator: One way to estimate $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$ is to replace μ_1 by its maximum likelihood estimator $\bar{X}$. In that case the estimator of $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$ is given by

$$(17) \qquad \hat{P}_1(\psi_0|\bar{X}_1, \bar{X}_2) = \Phi[(\bar{X}_1 + \bar{X}_2 - 2\bar{X}_1)'(\bar{X}_2 - \bar{X}_1)/2d]$$

$$= \Phi(d/2) \,.$$

The estimator given by (17) was originally proposed by Fisher [5] as an estimator of $P_1(\psi_0^*) = \Phi(\Delta/2)$ and as a natural choice. Though, given $\bar{X}_1$ and $\bar{X}_2$, this estimator is a constant, it is easy to find the unconditional distribution of the estimator. For $(1/2) \le t \le 1$ and $p = 1$, and for $\Phi^{-1}(\Phi(a)) = a$

$$(18) \qquad P[\Phi(d/2) \le t] = P[d \le 2\Phi^{-1}(t)]$$

$$= P[-2\Phi^{-1}(t) \le (\bar{X}_1 - \bar{X}_2) \le 2\Phi^{-1}(t)]$$

$$= \Phi\left[\frac{2\Phi^{-1}(t) - \mu_1+\mu_2}{\sqrt{\frac{2(1-\rho)}{N}}}\right] - \Phi\left[\frac{-2\Phi^{-1}(t) - \mu_1+\mu_2}{\sqrt{\frac{2(1-\rho)}{N}}}\right].$$

Let χ denote a noncentral chi-square random variable with p degrees of freedom and noncentrality $\Delta^2/2$. When $p > 1$, $d^2 \equiv \chi$ and hence

$$(19) \qquad P[\Phi(d/2) \le t] = P[x \le \frac{2(1-\rho)}{N} \{2\Phi^{-1}(t)\}^2] .$$

One might refer to [3] for exact expression of (19). When $p = 1$, the expected value of the estimator $\Phi(d/2)$ is obtained as, for $Z \sim N(0,1)$,

$$(20) \qquad E\Phi(d/2) = EP[Z \le d/2 | \bar{X}_1, \bar{X}_2]$$

$$= P[Z - (\bar{X}_2 - \bar{X}_1)/2 \le 0] + P[(\bar{X}_2 - \bar{X}_1) \le 0]$$

$$- 2P[Z - (\bar{X}_2 - \bar{X}_1)/2 \le 0, \ (\bar{X}_2 - \bar{X}_1) \le 0]$$

$$= \Phi\left[\frac{\Delta/2}{\sqrt{1 + \frac{1}{2N} - \frac{\rho}{2N}}}\right] + \Phi\left[\frac{-\Delta/2}{\sqrt{\frac{1}{2N} - \frac{\rho}{2N}}}\right]$$

$$- 2G\left[\frac{\Delta/2}{\sqrt{1 + \frac{1}{2N} - \frac{\rho}{2N}}}, \frac{-\Delta/2}{\sqrt{\frac{1}{2N} - \frac{\rho}{2N}}}, \sqrt{\frac{1-\rho}{2N+1-\rho}} \right]$$

Where $G[\cdot, \cdot, \rho]$ is the distribution function of standard bivariate normal with coefficient of correlation ρ. For $p > 1$, exact expression becomes complicated. However, for large N, one has (see [3])

$$(21) \qquad E\,\Phi(d/2) = \Phi(\Delta/2) + N^{-1}\varphi(\Delta/2)[(2\Delta)^{-1}(p-1)(1-\rho) - \Delta(1-\rho)/4] + O(N^{-2})$$

where $\varphi(x)$ is standard normal density function.

(b) Reclassification estimator: The reclassification estimator P_R, the proportion of observations from π_1 correctly classified by the rule ψ_0, was proposed by Smith [13] to estimate $P_1(\psi_0^*)$ and used by Hills [7] and Sorum [14] as an estimator of $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$. Results and expressions of Sorum [14] will be used here in studying P_R.

Let

$$Y_\alpha = \begin{cases} 1 & \text{if } ((X_{1\alpha} - \bar{X}_1)) < ((X_{1\alpha} - \bar{X}_2)) \\ 0 & \text{otherwise .} \end{cases}$$

Then

$$P_R = (1/N) \sum_{\alpha=1}^{N} Y_\alpha .$$

Now

$$(22) \qquad E(P_R|\bar{X}_1, \bar{X}_2) = E(Y_1|\bar{X}_1, \bar{X}_2) = Q_1, \text{ say}$$

and

$$(23)\qquad \mathrm{Var}(P_R|\bar{X}_1, \bar{X}_2) = N^{-1}\,\mathrm{Var}(Y_1|\bar{X}_1, \bar{X}_2) + (N-1)N^{-1}\,\mathrm{Cov}(Y_1, Y_2|\bar{X}_1, \bar{X}_2)$$

$$= N^{-1}(Q_1 - Q_1^2) + (N-1)\,N^{-1}(Q_{1,2} - Q_1^2)$$

$$= N^{-1}\,Q_1 + (N-1)\,N^{-1}\,Q_{1,2} - Q_1^2$$

where

$$(24)\qquad Q_{1,2} = E(Y_1 \cdot Y_2|\bar{X}_1, \bar{X}_2)\ .$$

Now

$$(X_{11}', X_{12}', \bar{X}_1', \bar{X}_2')' \sim N_{4p}\,[\lambda, \Lambda]$$

where

$$\lambda' = (\mu_1', \mu_1', \mu_1', \mu_2')$$

and

$$(25)\qquad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

and where Λ_{ij}'s are $2p \times 2p$ matrices given by

$$\Lambda_{11} = \begin{pmatrix} I_p & 0 \\ 0 & I_p \end{pmatrix}$$

$$\Lambda_{12} = \Lambda_{21}' = \begin{pmatrix} N^{-1} I_p & \rho N^{-1} I_p \\ N^{-1} I_p & \rho N^{-1} I_p \end{pmatrix}$$

and

$$\Lambda_{22} = \begin{pmatrix} N^{-1} I_p & \rho N^{-1} I_p \\ \rho N^{-1} I_p & N^{-1} I_p \end{pmatrix} .$$

Hence the conditional joint distribution of X_{11} and X_{12} given $\bar{X}_1$ and $\bar{X}_2$ is

$$(26) \qquad \left[\begin{pmatrix} X_{11} \\ X_{12} \end{pmatrix} \middle| \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} \right] \sim N_{2p} \left[\mu_c , \Lambda_{11 \cdot 2} \right]$$

where

$$(27) \qquad \mu_c = \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix} + \Lambda_{12} \Lambda_{22}^{-1} \begin{pmatrix} \bar{X}_1 - \mu_1 \\ \bar{X}_2 - \mu_2 \end{pmatrix}$$

$$= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \frac{1}{1-\rho^2} \begin{pmatrix} I_p & \rho I_p \\ I_p & \rho I_p \end{pmatrix} \begin{pmatrix} I_p & -\rho I_p \\ -\rho I_p & I_\rho \end{pmatrix} \begin{pmatrix} \bar{X}_1 - \mu_1 \\ \bar{X}_2 - \mu_2 \end{pmatrix}$$

$$= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} I_p & 0 \\ I_p & 0 \end{pmatrix} \begin{pmatrix} \bar{X}_1 - \mu_1 \\ \bar{X}_2 - \mu_2 \end{pmatrix} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix}$$

and

$$(28) \qquad \Lambda_{11 \cdot 2} = \Lambda_{11} - \Lambda_{12} \Lambda_{22}^{-1} \Lambda_{21}$$

$$= \Lambda_{11} - \begin{pmatrix} I_p & 0 \\ I_p & 0 \end{pmatrix} \begin{pmatrix} N^{-1} I_p & N^{-1} I_p \\ \rho N^{-1} I_p & \rho N^{-1} I_p \end{pmatrix}$$

$$= \begin{pmatrix} I_p & 0 \\ 0 & I_p \end{pmatrix} - \begin{pmatrix} N^{-1} I_p & N^{-1} I_p \\ N^{-1} I_p & N^{-1} I_p \end{pmatrix}$$

$$= \begin{pmatrix} (1 - \frac{1}{N})I_p & - N^{-1} I_p \\ - N^{-1} I_p & (1 - \frac{1}{N})I_p \end{pmatrix} .$$

Sorum [14, appendix II] has obtained the same result (26) with μ_c given by (27) and $\Lambda_{11\cdot 2}$ given by (28). Hence we have

$$(29) \qquad Q_1 = \Phi \left[\frac{(1/2)(\bar{X}_1 + \bar{X}_2)'(\bar{X}_2 - \bar{X}_1) - (\bar{X}_1)'(\bar{X}_2 - \bar{X}_1)}{\sqrt{(1 - \frac{1}{N})\, d^2}} \right]$$

$$= \Phi \left[\frac{d}{2} (1 - \frac{1}{N})^{-1/2} \right]$$

$$Q_{1,2} = P[((X_{11} - \bar{X}_1)) < ((X_{11} - \bar{X}_2)),\ ((X_{12} - \bar{X}_1)) < ((X_{12} - \bar{X}_2))\ \ \bar{X}_1, \bar{X}_2]$$

(30)

$$= G[(1- \frac{1}{N})^{-1/2} \frac{d}{2},\ (1- \frac{1}{N})^{-1/2} \frac{d}{2},\ - (N-1)^{-1}] .$$

Hence

$$E(P_R|\bar{X}_1, \bar{X}_2) = \Phi[\frac{d}{2} (1- \frac{1}{N})^{-1/2}]$$

and

$$V(P_R|\bar{X}_1, \bar{X}_2) = N^{-1} \Phi[\frac{d}{2} (1- \frac{1}{N})^{-1/2}]$$

$$+ (N-1)N^{-1} G[\frac{d}{2} (1- \frac{1}{N})^{-1/2}, \frac{d}{2} (1- \frac{1}{N})^{-1/2}, - (N-1)^{-1}]$$

$$- \Phi^2 \left[\frac{d}{2}\left(1 - \frac{1}{N}\right)^{-1/2}\right]$$

as obtained by Sorum. One inequality, which is obvious from (17) and (29), is

$$\hat{P}_1(\psi_0|\bar{X}_1, \bar{X}_2) < E(P_R|\bar{X}_1, \bar{X}_2) .$$

Since $E(P_R|\bar{X}_1, \bar{X}_2)$ and $\hat{P}_1(\psi_0|\bar{X}_1, \bar{X}_2)$ are similar in nature (except for the constant factor $(1 - \frac{1}{N})^{-1/2}$) we can obtain the distribution of $E(P_R|\bar{X}_1, \bar{X}_2)$ along the same line as was done for $\hat{P}_1$ and in particular, for $p = 1$

$$(31) \qquad E[E(P_R|\bar{X}_1, \bar{X}_2)] = E(P_R) = \Phi\left[\frac{\Delta/2}{\sqrt{1 - \frac{1}{2N} - \frac{\rho}{2N}}}\right] + \Phi\left[\frac{-\Delta/2}{\sqrt{\frac{1}{2N} - \frac{\rho}{2N}}}\right]$$

$$- 2G\left[\frac{\Delta/2}{\sqrt{1 - \frac{1}{2N} - \frac{\rho}{2N}}}, \frac{-\Delta/2}{\sqrt{\frac{1}{2N} - \frac{\rho}{2N}}}, -\sqrt{\frac{1-\rho}{2N-1-\rho}}\right] .$$

For $p > 1$ and for large N (see [3])

$$(32) \qquad E(P_R) = \Phi(\Delta/2) + N^{-1}\varphi(\Delta/2)[(2\Delta)^{-1}(p-1)(1-\rho) + \rho/4] + O(N^{-2}) .$$

(c) Lachenbruch's estimator P_U: The estimator P_U was introduced by Lachenbruch [8] which is defined as follows:

Let

$$\bar{X}_1^{(k)} = (N-1)^{-1} \sum_{\substack{\alpha=1 \\ \neq k}}^{N} X_{1\alpha} \qquad k = 1,2,\ldots,N \ .$$

Define Y_α as before, replacing $\bar{X}_1$ by $\bar{X}_1^{(\alpha)}$ and

$$P_U = (1/N) \sum_{\alpha=1}^{N} Y_\alpha \ .$$

This estimator was studied by Lachenbruch [8] and Lachenbruch and Mickey [9] as an estimator of $P_1(\psi_0^*)$. Hills [7] and Sorum [14] studied this as an estimator of $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$. Question may be raised about the appropriateness of using this estimator in the conditional sense since we are using different rules to classify each observation. A partial answer may be found in C.A.B. Smith's discussion of Hills [7] paper and Hill's reply, or in Sorum [13]. One justification for using P_U was that the observations to be classified were independent of the rest of the observations and so the computations were not much involved; but, in our case, since $X_{1\alpha}$ and $X_{2\alpha}$ are dependent, the classification statistic and the observation to be classified become dependent. However we shall see later that we get the same results obtained by Sorum [14] for $\rho = 0$. The estimator P_U will be studied only for $p = 1$.

Now,

$$\begin{pmatrix} \bar{X}_1^{(1)} \\ \bar{X}_1^{(2)} \\ \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} \sim N_4 \left[\begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \end{pmatrix} , \Lambda \right]$$

where

$$\Lambda = \begin{pmatrix} \frac{1}{N-1} & \frac{N-2}{(N-1)^2} & \frac{1}{N} & \frac{\rho}{N} \\ \frac{N-2}{(N-1)^2} & \frac{1}{N-1} & \frac{1}{N} & \frac{\rho}{N} \\ \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \frac{\rho}{N} \\ \frac{\rho}{N} & \frac{\rho}{N} & \frac{\rho}{N} & \frac{1}{N} \end{pmatrix}$$

Hence the conditional distribution of $\bar{X}_1^{(1)}$ and $\bar{X}_2^{(1)}$ given $\bar{X}_1$ and $\bar{X}_2$ is bivariate normal with conditional mean μ_c given by

$$\mu_c = \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix} + \begin{pmatrix} N^{-1} & \rho N^{-1} \\ N^{-1} & \rho N^{-1} \end{pmatrix} \begin{pmatrix} N^{-1} & \rho N^{-1} \\ \rho N^{-1} & N^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \bar{X}_1 - \mu_1 \\ \bar{X}_2 - \mu_2 \end{pmatrix}$$

$$= \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \bar{X}_1 - \mu_1 \\ \bar{X}_2 - \mu_2 \end{pmatrix} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_1 \end{pmatrix}$$

and the conditional covariance matrix $\Lambda_{11\cdot 2}$

$$\Lambda_{11\cdot 2} = \begin{pmatrix} (N-1)^{-1} & (N-2)(N-1)^{-2} \\ (N-2)(N-1)^{-2} & (N-1)^{-1} \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} N^{-1} & N^{-1} \\ \rho N^{-1} & \rho N^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} 1/N(N-1) & -1/N(N-1)^2 \\ -1/N(N-1)^2 & 1/N(N-1) \end{pmatrix}$$

Hence

$$E(Y_1 | \bar{X}_1, \bar{X}_2) = 1 - \Phi(-d/a) - \Phi(-d/b) \tag{33}$$

and

$$(34)\qquad E(Y_1\,Y_2|\bar{X}_1,\,\bar{X}_2)$$

$$= E\{(1-Y_1)(1-Y_2)|\bar{X}_1,\,\bar{X}_2\} + 2E(Y_1|\bar{X}_1,\,\bar{X}_2) - 1$$

$$= 1 - 2\Phi(-d/a) - 2\Phi(-d/b) + G[\tfrac{d}{a},\,\tfrac{d}{a},\,-\tfrac{1}{N-1}]$$

$$- 2G[\tfrac{d}{a},\,\tfrac{d}{b},\,\tfrac{1}{N-1}] + G[\tfrac{d}{b},\,\tfrac{d}{b},\,-\tfrac{1}{N-1}]$$

where $b = N^{-1/2}(N-1)^{-1/2}$ and $a = (2N-1)b$. Hence the conditional variance is easily calculated from (33), (34) and (23). The unconditional expected value is given by

$$(35)\quad 1 - EP_U = P[2X_{11} - \bar{X}_1^{(1)} - \bar{X}_2 \le 0] + P[\bar{X}_1^{(1)} - \bar{X}_2 \le 0]$$

$$- 2P[2X_{11} - \bar{X}_1^{(1)} - \bar{X}_2 \le 0,\ \bar{X}_1^{(1)} - \bar{X}_2 \le 0]$$

$$= \Phi\left(\frac{\Delta/2}{\sqrt{1+e}}\right) + \Phi\left(\frac{\Delta/2}{\sqrt{e}}\right) - 2G\left[\frac{\Delta/2}{\sqrt{1+e}},\ \frac{\Delta/2}{\sqrt{e}},\ -\frac{2(N-1)\rho+1}{\sqrt{(4N^2-4N+f)f}}\right]$$

where

$$e = [2N - 1 - 2\rho(N-1)]/4N(N-1) = f/4N(N-1)\ .$$

(d) Estimators using prior distribution: Various estimators are derived by Sorum [14] using a normal prior for the unknown mean vectors. The simplest estimator using a prior distribution is

obtained by replacing the unknown mean vector by its conditional posterior mean given the training sample. Another estimator is obtained by considering the conditional expected value of the posterior distribution of $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$ given the training sample. The later method was suggested by Geisser [6]. If a quadratic loss function is considered, estimators given above minimize the posterior (conditional) expected loss (see Sorum [14]). Even though $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$ does not involve μ_2, it might be reasonable to assume correlated prior distributions for μ_1 and μ_2 since populations π_1 and π_2 are correlated. In particular, the prior distribution of μ_1 and μ_2 is assumed to have a $2p$ variate normal distribution given by

$$(36) \qquad \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N_{2p} \left(\begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} , \begin{pmatrix} a^2 I_p & \rho a^2 I_p \\ \rho a^2 I_p & a^2 I_p \end{pmatrix} \right) .$$

The reason for using the same ρ is because, as was in the example,

$$E(X_t) = m_t + \lambda E(X_{t-1})$$

i.e., μ_t's could be assumed to have a first order auto correlated model. The limiting form of the estimators using the prior distribution given by (36) as $a^2 \to \infty$ may be considered as no prior information.

Straight forward calculations show that if the distribution of Z_1 given Z_2 is $N_p[Z_2, \Lambda_1]$ and the distribution of Z_2 is $N_p[\mu, \Lambda_2]$, then the distribution of Z_2 given Z_1 is

$$(37) \qquad N_p \left[(\Lambda_1^{-1} + \Lambda_2^{-1})^{-1} (\Lambda_1^{-1} Z_1 + \Lambda_2^{-1} \mu), (\Lambda_1^{-1} + \Lambda_2^{-1})^{-1} \right] .$$

Hence, the conditional posterior distribution of μ_1 and μ_2 given $\bar{X}_1$ and $\bar{X}_2$ is a 2p-variate normal with covariance matrix given by

$$(38)\quad \left\{ \begin{pmatrix} N^{-1}I_p & N^{-1}\rho I_p \\ N^{-1}\rho I_p & N^{-1}I_p \end{pmatrix}^{-1} + \begin{pmatrix} a^2 I_p & \rho a^2 I_p \\ \rho a^2 I_p & a^2 I_p \end{pmatrix}^{-1} \right\}^{-1}$$

$$= (N + a^{-2})^{-1} \begin{pmatrix} I_p & \rho I_p \\ \rho I_p & I_p \end{pmatrix}$$

and the mean vector is

$$(N + a^{-2})^{-1} \begin{pmatrix} I_p & \rho I_p \\ \rho I_p & I_p \end{pmatrix} \left\{ \begin{pmatrix} \frac{N}{1-\rho^2} I_p & -\frac{N}{1-\rho^2}\rho I_p \\ -\frac{N}{1-\rho^2}\rho I_p & \frac{N}{1-\rho^2} I_p \end{pmatrix} \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} \right.$$

$$\left. + \begin{pmatrix} \frac{1}{a^2(1-\rho^2)} I_p & -\frac{\rho}{a^2(1-\rho^2)} I_p \\ -\frac{\rho}{a^2(1-\rho^2)} I_p & \frac{1}{a^2(1-\rho^2)} I_p \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \right\}.$$

This reduces to

$$(39)\quad \left(\frac{a^2 N \bar{X}_1 + \delta_1}{1 + a^2 N} \quad \frac{a^2 N \bar{X}_2 + \delta_2}{1 + a^2 N} \right)' .$$

Hence the marginal conditional posterior distribution of μ_1 given $\bar{X}_1$ and $\bar{X}_2$ is

$$N_p \left[\frac{a^2 N \bar{X}_1 + \delta_1}{1 + a^2 N}, \; (N + a^{-2})^{-1} I_p \right].$$

Hence an estimator of $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$ is

$$(40) \qquad \Phi\left[\left(\bar{X}_1 + \bar{X}_2 - 2\,\frac{a^2 N \bar{X}_1 + \delta_1}{1 + a^2 N}\right)'(\bar{X}_2 - \bar{X}_1)/2d\right]$$

which is identical to Fisher's estimator $\Phi(d/2)$ as $a^2 \to \infty$.

The other estimator is derived by considering the conditional posterior mean of $P_1(\psi_0|\bar{X}_1, \bar{X}_2)$. We note that for

$$\theta = \left(\bar{X}_1 + \bar{X}_2 - 2\,\frac{a^2 N \bar{X}_1 + \delta_1}{1 + a^2 N}\right)'(\bar{X}_2 - \bar{X}_1)/2d$$

and $\alpha^2 = (N + a^{-2})^{-1}$ we have the conditional posterior distribution of

$$T \equiv (\bar{X}_1 + \bar{X}_2 - 2\,\mu_1)'(\bar{X}_2 - \bar{X}_1)/2d \sim N[\theta, \alpha^2] .$$

Hence

$$(41) \qquad E_{\mu_1|\bar{X}_1,\bar{X}_2} P_1(\psi_0|\bar{X}_1, \bar{X}_2)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{t} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\frac{e^{-(t-\theta)/2\alpha^2}}{\sqrt{2\pi}\,\alpha}\,dx\,dt$$

$$= \Phi\left[\theta/\sqrt{1 + \alpha^2}\right]$$

which, when $a^2 \to \infty$, converges to the estimator given by

$$(42) \qquad \Phi\left(\frac{d}{2}/\sqrt{\left(1 + \frac{1}{N}\right)}\right) .$$

Both of these estimators (40) and (41) agree with the estimators obtained by Sorum [14]. The unconditional expectation or the distribution of (42) can be derived in the similar line as was done for Fisher's estimator.

4. A more general set-up.

Consider a situation when $\{X_t\}$ is a stationary Gaussian process and the two populations are any two time points. Then (see Anderson [2], page 173) $(X'_{1\alpha}, X'_{2\alpha})'$, $\alpha = 1,2,\ldots,N$ are independently distributed as

$$(43) \qquad N_{2p}\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \Gamma\right]$$

where

$$(44) \qquad \Gamma = \begin{pmatrix}\Sigma & \tau\\ \tau & \Sigma\end{pmatrix}$$

and where Σ is a $p \times p$ positive definite matrix and τ is a $p \times p$ matrix such that Γ is a $2p \times 2p$ positive definite matrix. We note that (1) is a special case of (43) when $\tau = \rho\Sigma$ or $p = 1$. We first establish that the rule ψ_k is the likelihood ratio rule in this case. The supremum over μ_1 and μ_2 of the likelihood of the training sample and X, when X is from π_i, is given by

$$(45) \qquad L_i(\tau) = (2\pi)^{-p(2N+1)/2}|\Sigma|^{-1/2}$$
$$\cdot \exp\{\text{trace}[-(1/2)\Sigma^{-1} D_{0i}]\}$$
$$\cdot |\Gamma|^{-N/2} \exp\{\text{trace}[-(1/2)\Gamma^{-1} S]\}$$

where

$$(46) \qquad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

and S_{ij}'s are given by (12) and D_{0i} is given by (14). Since, for τ known or unknown

$$L_1(\tau)/L_2(\tau) = \sup_{\tau} L_1(\tau) / \sup_{\tau} L_2(\tau)$$

ψ_k is the likelihood ratio rule. This rule is also the plug-in rule and the minimum distance rule, but, unlike the case when $\tau = \rho \Sigma$, we have not established the admissibility of ψ_k for general τ.

Results for Fisher's estimator $\Phi(d/2)$ are still true except that the large sample result given in (21) changes to (see [3])

$$(47) \quad E\Phi(d/2) = \Phi(\Delta/2) + N^{-1} \varphi(\Delta/2)[(2\Delta)^{-1} \text{ trace } (I_p - \Sigma^{-1} \tau) - \{(2\Delta^3)^{-1} + (4\Delta)^{-1}\} (\mu_1-\mu_2)'(\Sigma^{-1} - \Sigma^{-1} \tau \Sigma^{-1})(\mu_1-\mu_2)] + O(N^{-2}) .$$

Similar computations show that for Smith's estimator P_R, (32) changes to

$$(48) \quad E[P_R] = \Phi(\Delta/2) + N^{-1} \varphi(\Delta/2) [(2\Delta)^{-1} \text{ trace } (I_p - \Sigma^{-1} \tau) - \{(2\Delta^3)^{-1} + (4\Delta)^{-1}\}(\mu_1-\mu_2)'(\Sigma^{-1} - \Sigma^{-1} \tau \Sigma^{-1})(\mu_1-\mu_2) + \Delta/4] + O(N^{-2}) .$$

However, conditional distribution of P_R remains the same and hence the conditional expectation and conditional variance are unchanged. This fact can be seen from the following analysis.

Note that, there exists a matrix A such that (see Rao [12], page 37)

$$A \Sigma A' = I_p \tag{49}$$

and

$$A \tau A' = D \tag{50}$$

where D is a diagonal matrix with diagonal elements less than unity in absolute value (since Γ is positive definite). Moreover, the class of rules ψ_k^* and ψ_k are invariant under the transformations

$$\begin{aligned} &X_{i\alpha} \to A X_{i\alpha} \quad , \quad i = 1,2; \quad \alpha = 1,2,\ldots,N \\ &X \to A X \end{aligned} \tag{51}$$

for every nonsingular matrix A. Hence without loss of generality we can take $\Sigma = I_p$ and $\tau = D$, a diagonal matrix. Now, we observe that Λ in (25) is given by

$$\Lambda_{11} = \begin{pmatrix} I_p & 0 \\ 0 & I_p \end{pmatrix} \quad , \quad \Lambda_{22} = N^{-1} \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix}$$

and

$$\Lambda_{12} = \Lambda_{21}' = N^{-1} \begin{pmatrix} I_p & D \\ I_p & D \end{pmatrix}.$$

Hence (see Rao [12], page 29)

$$\Lambda_{12}\,\Lambda_{22}^{-1} = \begin{pmatrix} I_p & D \\ I_p & D \end{pmatrix} \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} I_p & D \\ I_p & D \end{pmatrix} \begin{pmatrix} I_p + D(I_p - D^2)^{-1} D & -D(I_p - D^2)^{-1} \\ -D(I_p - D^2)^{-1} & (I_p - D^2)^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} I_p & 0 \\ I_p & 0 \end{pmatrix}$$

Hence (27) and (28) still hold.

To see the estimators derived using prior distributions on μ_1 and μ_2 remains unchanged, let the joint prior distribution of μ_1 and μ_2 be, after reduction of the problem,

$$(52) \qquad \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N_{2p}\left[\begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}, a^2 \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix} \right] .$$

Once again using (37), we get the conditional covariance matrix of μ_1 and μ_2 given $\bar{X}_1$ and $\bar{X}_2$ as

$$\left\{ N \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix}^{-1} + a^{-2} \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix}^{-1} \right\}^{-1}$$

$$= (N + a^{-2})^{-1} \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix}$$

and so, the conditional mean vector of μ_1 and μ_2 given $\bar{X}_1$ and $\bar{X}_2$ is

$$(N + a^{-2})^{-1} \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix} \left\{ N \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix}^{-1} \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} + a^{-2} \begin{pmatrix} I_p & D \\ D & I_p \end{pmatrix}^{-1} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \right\}$$

$$= (N + a^{-2})^{-1} \begin{pmatrix} N\bar{X}_1 + a^{-2}\delta_1 \\ N\bar{X}_2 + a^{-2}\delta_2 \end{pmatrix}$$

which is the same as (39).

Conditional asymptotic expansions for large N, given $\bar{X}_1$ and $\bar{X}_2$, of $\Phi(d/2)$, P_R and P_U (for $p = 1$) are obtained by Sorum [14], [16], and [17]. For the two Bayes estimators Sorum's results for $\tau = 0$ can be easily modified to accomodate our case.

Exact and large sample unconditional expectation of the conditional PCC has been considered in [3] along with the distribution of the associated classification statistic for unknown Σ. For $\tau = 0$ McLachhan [10] and [11] derived the asymptotic expansion of the unconditional distribution of the conditional probability of misclassification and, in particular, he has obtained the expansions of the unconditional expectation and the unconditional variance.

REFERENCES

[1] Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

[2] Anderson, T. W. (1971). The Statistical Analysis of Time Series. Wiley, New York.

[3] Bandyopadhyay, S. (1974). Classification with dependent training samples. Ph.D. Thesis, University of Minnesota.

[4] Ellison, B. E. (1965). Multivariate normal classification with covariance known. Ann. Math. Statist. 36 1787-1793.

[5] Fisher, R. A. (1936). Use of multiple measurements in Taxonomy problems. Ann. Eugen. 7 179-188.

[6] Geisser, S. (1967). Estimation associated with linear discriminants. Ann. Math. Statist. 38 807-817.

[7] Hills, M. (1966). Allocation rules and their error rates. J. Roy. Statist. Soc. Ser. B 28 1 - 32.

[8] Lachenbruch, P. A. (1965). Estimation of error rates in discriminant analysis. Ph.D. Thesis, University of California, Los Angeles.

[9] Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics 14 217-226.

[10] McLachlan, G. J. (1972). An asymptotic expansion for the variance of the errors in misclassification of the linear discriminant function. Aust. J. Statist. 14 68-72.

[11] McLachlan, G. J. (1974). The asymptotic distributions of the conditional error rate and risk in discriminant analysis, Biometrika 61 131-135.

[12] Rao, C. R. (1965). Linear Statistical Inference and Its Applications. Wiley, New York.

[13] Smith, C. A. B. (1947). Some examples in discrimination. Ann. Eugen. 13 272-282.

[14] Sorum, M. J. (1968). Estimating the probability of misclassification. Technical Report No. 110, University of Minnesota.

[15] Sorum, M. J. (1971). Estimating the conditional probability of misclassification. Technometrics 13 333-343.

[16] Sorum, M. J. (1972). Three probabilities of misclassification. Technometrics 14 309-316.

[17] Sorum, M. J. (1972). Estimating the expected and the optimal probabilities of misclassification. Technometrics 14 935-943.